PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-022414
(43)Date of publication of application : 21.01.1997

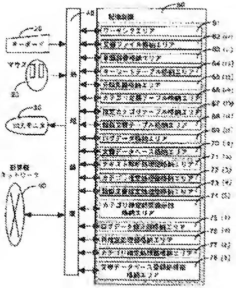
(51)Int.Cl. G06F 17/30

(21)Application number : 07-170682
(22)Date of filing : 06.07.1995

(71)Applicant : HITACHI LTD
(72)Inventor : MASE HISAO
MORIMOTO YUKIKO
TSUJI HIROSHI

(54) DOCUMENT SORTING SUPPORTING METHOD AND ITS DEVICE

(57)Abstract:
PROBLEM TO BE SOLVED: To shorten a checking time by collecting together the documents of similar contents based on an estimated category and showing these documents in sequence to a user to urge him to check them.
SOLUTION: The necessary hardware is composed of a keyboard 20 which inputs the operating instructions and the data given from a user, a mouse 25, an output monitor 30 which outputs the results, a processor 40 which carries out various types of processing, and a storage 50 which stores the files and programs. These hardware are connected to a computer network 90 and can acquire the document data via the network 90. When the sorting results of a computer are checked by the user, the log data are displayed to the user to show the relation between the document to be sorted and its corresponding category. Then the log data are corrected by the user, and a category is estimated again based on the corrected log data. The estimated category is displayed to the user.



CLAIMS

[Claim(s)]
[Claim 1]A document group support method using an input device, an output unit, and a processing unit that has memory storage, a) Analyze text information which stores in said memory storage a document including text information, and is included in a document by which the b aforementioned input was carried out, c) Presume a category of said inputted document using classification knowledge beforehand defined as said text-analysis result, d) Said presumed category authorizes similar or same document set as a similar document mutually, e) A document group support method becoming final and conclusive a category which should be classified based on directions which presumed a category which should be classified about said two or more similar documents, respectively, and were inputted from said input device about the f aforementioned similar document according to said estimation result displayed on said output unit.
[Claim 2]The document group support method according to claim 1 being able to choose whether a presenting method of outputting said estimation result to said output unit one by one is adopted in said step f via said input device.
[Claim 3]The document group support method according to claim 1 authorizing a document set which extracts words and phrases contained in the document about said two or more documents in said step d, respectively, and has said extracted words and phrases in common as a similar document.
[Claim 4]The document group support method according to claim 3 choosing whether a presenting method of showing said output unit said similar document one by one is adopted via said input device.
[Claim 5]The document group support method according to claim 1 specifying from which document set it shows via said input device in said step d when showing said output unit said document set.
[Claim 6]A document group support method using an input device, an output unit, and a processing unit that has memory storage, a) Analyze text information which stores in said memory storage a document including text information, and is included in a document by which the b aforementioned input was carried out, c) Presume a category of said inputted document using classification knowledge beforehand defined as said text-analysis result, d) Said presumed category authorizes similar or same document set as a similar document mutually, e) Presume a category which should be classified about said two or more similar documents, respectively, f) Based on directions inputted from said input device about said similar document according to said estimation result displayed on said output unit, becoming final and conclusive a category which should be classified -- g -- log data about why the document concerned was classified into

the category concerned being shown to said output unit, and. h) correcting said shown log data via said input device -- i -- re-presuming a category based on log data after said correction -- j -- a document group support method showing said output unit said re-presumed category.

[Claim 7]The document group support method according to claim 6 showing said output unit in a different mode from other categories about a category which does not exist in an estimation result before re-presuming said re-presumed result in said step j as compared with an estimation result before re-presuming.

[Claim 8]In said step g, to log data about why said document was classified into the category concerned. The document data concerned, phrase data extracted from a text contained in the document concerned, Phrase data by which each category defined as said classification knowledge is characterized, The document group support method according to claim 6 containing phrase-correspondences data about whether it is contained in words and phrases by which each category in which words and phrases extracted from a text contained in the document concerned are defined as classification knowledge is characterized, and category definition data which defined the range of each category.

[Claim 9]The document group support method according to claim 8, wherein words and phrases extracted from said text and words and phrases by which said each category is characterized have the dignity which shows importance of the words and phrases, respectively.

[Claim 10]The document group support method according to claim 6 deleting, adding and correcting some of these data via said input device about said document data, phrase data extracted from a text contained in a document, and phrase data by which each category defined as classification knowledge is characterized.

[Claim 11]The document group support method according to claim 9 correcting dignity which shows importance of words and phrases by which dignity or each category which shows importance of words and phrases extracted from said text is characterized via said input device.

[Claim 12]The document group support method according to claim 8 presuming a category in having limited to a subordinate category set which displays a category set which consists of multiple layers, makes a certain high order hierarchy's category specify via said input device, and belongs to the specified high order hierarchy category concerned.

[Claim 13]Claim 1 displaying definite category information and log data about a document which have been category become final and conclusive on said output unit before one or more documents, or the document group support method according to claim 6.

[Claim 14]a) A document input means which inputs a document including text information, a text analyzing means which analyzes text information included in a document by which the b aforementioned input was carried out, c) A category estimation means which presumes a category of said inputted document using classification knowledge beforehand defined as said text-analysis result, d) A similar document authorization means by which said presumed category authorizes similar or same document set as a similar document mutually, e) A similar document category estimation means which presumes a category which should be classified about said two or more similar documents, respectively, f) A document group support device having a category decision means to become final and conclusive a category which should be classified, based on directions inputted from said input device about said

similar document according to said estimation result displayed on said output unit.

[Translation done.]

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Industrial Application]This invention relates to the document group support method and device for doing efficiently the work in which a user checks especially an electronic document including text information to the classification result by a computer about the document sorting method and device which are classified into a category.
[0002]
[Description of the Prior Art]It is indispensable for a lot of information to come to overflow and to take out required information efficiently with social computerization and maintenance of the information infrastructure. Classifying a document into a suitable category beforehand is mentioned to one of solutions, and development of the automatic-classification art by a computer has been required.
[0003]As automatic-classification art of an electronic text document, Proceedings of second AnnualConference on Innovative (1990), There is art indicated to Information Processing Society of Japan report-of-research NL-98-11, Info-Tech'94 lecture collected-papers pp.138 - pp.146. These determine a category based on the appearance tendency of the keyword in a text document.
[0004]
[Problem(s) to be Solved by the Invention]The above-mentioned art is full automatic, classifies a text according to a computer, and is not mentioned in above-mentioned document about the method of determining a classification result as a user cooperatively. The classification accuracy by the above-mentioned art has not resulted in human being and an equivalent level.
[0005]However, in the situation where the classification accuracy of human being and an equivalent level is required, a user needs to check the classification result of a computer. Therefore, it leads to cost reduction that a computer and a user do division of roles and perform classifying cooperatively. That is, it becomes SUBJECT how the category which should be classified according to few [that it is efficient and] work burdens is become final and conclusive based on the classification result of a computer.
[0006]When the number of the documents which are the targets of sorting processing especially is extensive, the working hours which per affair takes are made few [how], and it becomes SUBJECT how a work burden is eased. The work which judges whether the classification result which the computer outputted is the right when there are comparatively many categories, or when a category is complicated and the

discernment is very difficult, and the work which finds a true category from 1 when the classification result is an error become very difficult. Therefore, it becomes SUBJECT how these work is done efficiently.

[0007]Then, one purpose of this invention is to do efficiently the work which judges whether a classification result is the right, and the work which finds a true category when the classification result is an error.

[0008]When classifying a lot of documents one by one, it does not depend for the turn on the contents of the document in many cases. In that case, in order for the contents described whenever the document changed to change a lot, whenever the user who checks changes his contents, he needs to shift gears to the contents. For this reason, SUBJECT that the efficiency of a check is bad and a work burden also increases occurs.

[0009]Then, the contents of the document which should be classified ease the work burden by changing a lot frequently, and one purpose of everything but this invention has them in raising the efficiency of classifying.

[0010]

[Means for Solving the Problem]Log data about why a document which is the target of sorting processing was classified into the category according to this invention is shown to a user via an estimation result output means of a category, An aforementioned problem is solved by making a user correct shown log data via a user input means, re-presuming a category based on log data after correction, and showing a user a category after re-presumption via an estimation result output means.

[0011]In this invention, a category which should be classified according to a text analyzing means and a category estimation means about two or more documents is presumed, respectively, A presumed category is mutually similar or it has a similar document authorization means to authorize the same document set, About a similar document, an aforementioned problem is solved by making a category which should show a user one by one a category presumed by a category estimation means via an estimation result output means, and should classify it into a user via a user input means about a shown document become final and conclusive.

[0012]

[Function]Since the document with which the contents are similar is packed based on the presumed category, a user is shown one by one and a check is urged, the change of a user's head accompanying a big change of the contents ends few, and a work burden reduces. Since the document with which the contents were similar continues, it becomes easy to harness tips when a former document is checked, know how, teachings, data, etc. in the check of a next document, and check operation time ends few.

[0013]

[Example]Working example of this invention is hereafter described in detail using Drawings. This example classifies a newspaper article into a certain category, and stores it in a document data base. The newspaper article data stored in the database for every category can be searched by using a publicly known search system.

[0014]Drawing 1 is a figure showing the outline of this example. First, the document which is the target of a classification is inputted in the document input 1. Document data may be acquired from the exterior via a network, may be acquired via media, such as a floppy disk, and may be acquired via handwriting input devices, such as voice recognition equipment, an image recognition device (character recognition is included), and a pen, etc. Document data may be acquired collectively periodically and the

document data which is circulating may be acquired one by one irregularly. The acquired document data is temporarily stored in the document file 10.

[0015]Next, document data is analyzed with the execution instruction of specification of the document data classified from a user, and category presumption. It distinguishes (1a), and when there is nothing, he follows to Step 3a whether there is any document which is not presumed.

[0016]First, in a certain case, it is the text analysis 2, and it carries out automatic extracting of the keyword by which the contents are characterized by natural language processing from a text. That is, with reference to the word dictionary 11 which stored a word, and its part of speech and conjugation information, a text is divided into a word, and a part of speech makes a keyword the word which is a noun, and stores in the key word table 12 with the frequency of occurrence of each keyword.

[0017]next, category presumption -- with reference to the category definition table 14 which defined the system of the classification knowledge 13 and the category which defined and stored the keyword by which each category is characterized beforehand by 3, In which category the keyword of the key word table 12 extracted from the text is contained searches, and a score is given to the category when contained. And the high category of a score presumes that it is a category which should classify the text. An estimation result is stored in the presumed category table 15. The data of the keyword information used when presuming a category, the score information of a category, etc. is stored in the log data 17.

[0018]Next, an estimation result is outputted in order to make a user check an estimation result. A user is made to specify whether at this time, an estimation result is displayed for every document with which those contents are similar (3a), and when not displaying for every document with which the contents are similar, an estimation result is displayed in order of a document ID.

[0019]When displaying for every document with which the contents are similar, it is the similar document authorization 4, and from the category estimation result of each document stored in the presumed category table 15, a similar document is authorized and the result is stored in the similar document table 16.

[0020]Next, if there is a document in which the category is not become final and conclusive by the user (4a), decision of the category which should show a user a category estimation result one by one, and should check and classify (5) and a result will be urged (5a). The analytical data stored in the log data 17 at this time are also shown to a user.

[0021]The category shown the user confirms whether to be the right. And if it is that of the right, a category will be become final and conclusive and a document will be registered into the document data base 18. A right category must be found if not right. When it points represuming then the category which a user should classify, first, about the log data shown, a user is made to correct, based on the data after (6) and correction, a category is re-presumed and (7) and a new estimation result are shown to a user with new analytical data. When a user judges by this that it is a right category, a category is become final and conclusive and it registers with (8) and a document data base (9). Even if it performs re-presumption of a category several times, when a right category cannot be found, a user becomes final and conclusive a category with a help.

[0022]If a category is become final and conclusive, it will shift to the check of the following document (9a), and the category estimation result and log data of the document will be outputted.

[0023]Drawing 2 is a figure showing the outline of the hardware of this example. It consists of the keyboard 20 for inputting the operator guidance and data from a user, the mouse 25, the output monitor 30 that outputs a result, the processing unit 40 which performs various processings, and the memory storage 50 which stores a file and a program. In order to acquire document data, it is connected with the computer network 90 and a document can be acquired via a network.
[0024]The memory storage 50, Temporary data. The working area 61 to store and the acquired document data. The document file storage area 62, the word dictionary storage area 63, the key word table storage area 64, the classification knowledge storage area 65, the category definition table storage area 66, the presumed category table storage area 67, the similar document table storage area 68 which carry out a temporary storage, The log data storage area 69 and the document data base storage area 70 are included. The file of a data format is stored in the above-mentioned storage areas other than working area 61.
[0025]The memory storage 50, The text-analysis treating part storage area 71, the category estimation processing part storage area 72, the similar document authorization treating part storage area 73, the category estimation result indicator storage area 74, the log data corrected part storage area 75, the category re-estimation processing part storage area 76, the category determining processing part storage area 77, The document data base registration processing section storage area 78 is also included. The load module file of executable code is stored in these storage areas.
[0026]The number in () shown in drawing 2 shows a correspondence relation with each part shown in drawing 1.
[0027]Drawing 3 is a figure showing an example of the text information included in a document. Although the document data treated by this example is a newspaper article, as document data, the thing of other kinds, such as electronic news, an E-mail, a technology paper, a patent specification, claim and question / opinion sentence, and a conference note of a meeting, may be sufficient as it. In this example, these information is premised on being stored in a file in text code form to document data on the assumption that text information is included. However, that to which Still Picture Sub-Division, an animation, speech information, etc. are linked does not interfere.
[0028]Drawing 4 is a figure showing an example of the word dictionary 11 referred to by the text analysis 2. A word dictionary has word attribution information, such as the part of speech 202, the practical use kind 203, and the practical use line 204 besides the title 201.
[0029]Drawing 5 is a figure in the text analysis 2 showing an example of word division results. First, to a text like drawing 3, with reference to the word dictionary 11 of drawing 4, each sentence is divided for every word and the title 211 and the part of speech 212 of a word are extracted like drawing 5 in the text analysis 2. the concrete realization method of word division -- Information Processing Society of Japan 44th -- as shown in time national conference collected-papers (3)3-181, since it is publicly known, a detailed description is already omitted here.
[0030]Drawing 6 is a figure showing an example of the key word table 12 which stores the keyword extracted from the text. In the text analysis 2, after carrying out word division of the text, a part of speech extracts the word which is a noun, considers it as a keyword, computes the frequency of occurrence of each keyword in the text concerned further, and considers it as the dignity of a keyword. Of course, it is good considering parts of speech other than a noun as a keyword, and weighting of the

frequency of occurrence may be carried out in consideration of the appearing position of a keyword, a relation with the word before and behind that, etc. also besides considering it as dignity. The key word table 12 consists of the document ID 221 which identifies a document, the keyword title 222, and its dignity 223.
[0031]Drawing 7 is a figure showing an example of the category definition table 14 which defined the system of the category. This example defines the category which consists of two hierarchies called the large category 231 and the small category 232 as a category for classifying a newspaper article. The one or more small categories 232 belong to each of the large category 231, and the system of the tree structure is made it. the hierarchy of a category -- what floor layer -- it may be.
[0032]Drawing 8 is a figure showing an example of the classification knowledge 13. In this example, how to presume the category which should be classified based on the existence of a keyword is used. Therefore, the classification knowledge 13 is a set of the keyword by which a category is characterized. That is, the classification knowledge 13 consists of the large category 241, the small category 242, the keyword 243 by which the category is characterized, and the dignity 244 depending on the importance of the keyword. The dignity 244 has so large a value that the keyword is an important keyword by which the category is characterized. This classification knowledge 13 is beforehand stored in the memory storage 50. Classification knowledge may be created by a help and may be created by preparing the text which the category has already become final and conclusive according to a category, and carrying out automatic extracting of the keyword for every category.
[0033]drawing 9 -- category presumption -- it is a figure showing the procedure of 3. First, the table which stores the score of each category is initialized to 0 (Step 501).
[0034]Next, the following processings are performed about all the keywords of the document concerned stored in the key word table 12 (Step 502). Distinguish whether the category in the classification knowledge 13 containing the keyword concerned exists (Step 503), and about the existing category. The product of the dignity Wi (equivalent to 223 of drawing 6) which the keyword of the document concerned has, and the dignity Wj (equivalent to 244 of drawing 8) which the keyword of the category concerned has is calculated, and it adds as a score of the category concerned (Step 504).
[0035]Since it opts for the score of each category when the above-mentioned processing is performed about all the keywords, the deviation score of a score of each category is calculated from these scores (Step 505). A category is sorted in order with high deviation score (Step 506). And the document ID concerned, a category, and the value of the deviation score are made into a group, and it stores in the presumed category table 15 at order with high deviation score (Step 507). Top three categories are stored in this example. Of course, top n categories may be stored, a minimum may be provided in the value of deviation score and the category more than a minimum may be stored. Finally, when the document ID concerned, the keyword extracted from the document concerned, and each keyword are contained in the keyword of each category which it has, they store dignity Wi of Step 504, the dignity Wj, and the value of the product in the log data 17 (Step 508).
[0036]in addition -- although this example is making two hierarchies' (a large category, a small category) category system -- category presumption -- in 3, it carries out about a small category, and since it will be decided that it will be a meaning if a small category is decided, presumption of the large category is omitted. The method of presuming a category about a large category and presuming a small category first as

another estimation method in the form limited to the large category ranked as the higher rank may be used. In this case, the classification knowledge 13 which defined the keyword by which a large category is characterized, and its dignity is required. It may newly create by a help and can also create easily by arranging the classification knowledge about a small category for every large category.

[0037]Drawing 10 is a figure showing an example of the presumed category table 15. The presumed category table 15 consists of the document ID 251, the ranking 252 of the presumed category, the presumed large category candidate 253, the presumed small category candidate 254, and the deviation score 255 of the category.

[0038]Drawing 11 is a figure showing the procedure of the similar document authorization 4. First, the similar document table 16 is initialized (Step 521). Next, the following processings are performed about all the categories (Step 522). with reference to the presumed category table 15, it should classify into the category concerned at primacy in the document which presumed the category -- the document ID of the presumed document is extracted (Step 523).

[0039]next, it should classify into the 2nd place about the extracted document ID -- it collects for every presumed category, it matches with the category concerned, and stores in the similar document table 16 (Step 524).

[0040]Drawing 12 is a figure showing an example of the similar document table 16. As shown in drawing 11, in this example, the category presumed by primacy and the category presumed by the 2nd place are summarized for every same document, and stores in the similar document table 16. That is, the similar document table 16 comprises the category 261 presumed by primacy, the category 262 presumed by the 2nd place, and the document ID 263 which has them as an estimation result.

[0041]Drawing 13 is a figure showing an example of a category estimation result display. Here, the document designation button 401 specifies the range of the document to process.

The directory where a document exists is specified.

About the specified document, the classification button 402 performs the text analysis 2 and category presumption 3, and obtains an estimation result and log data. The reclassification button 403 performs re-presumption of a category based on the data corrected by the user, and outputs a re-estimation result. The narrowing-down classification button 404 makes a user specify a high order hierarchy's category, performs category presumption in having limited to the subordinate category belonging to the category, and outputs an estimation result so that it may mention later. The category list button 405 displays the contents of the category definition table 14. The classification knowledge reference button 406 displays the keyword stored in the classification knowledge 13, and its dignity according to a category. The end button 407 ends a system.

[0042]411 is area which displays the contents of the text.

ID of a document text is also displayed.

412 is area which makes a pair the keyword extracted from the text concerned, and its dignity (frequency of occurrence), and is displayed on order with high dignity.

[0043]Which keyword 413 containing among the keywords of 412 about each category and its score indicate what size it is. Specification of a category is performed by specifying any one of the categories of 414 which is a classification result. It is 413 of drawing 13 and the keyword a "circle" is contained in the keyword of the small category the "international economy", for example.

The dignity Wj which the keyword "circle" of the small category [dignity / Wi / which

the keyword extracted from the text has] the "international economy" of 4 and the classification knowledge 13 has shows that 8, as a result 8= 32 4x scores were given.

[0044]414 is the presumed large category, a small category, and area which displays the deviation score. 415 is area which displays the category which the user became final and conclusive. 416 is a button which displays the category estimation result and log data, and a definite category about the document checked just before the document checked now. Since the data about the document checked [these] is stored in the presumed category table and the log data, it is easily realizable by displaying those data.
[0045]417 is a button which directs to become final and conclusive a category about the document checked now, and to shift to the check of the following document. It becomes final and conclusive as a category which should classify the category described by 415 at this time, and the document concerned is registered into the document data base 18 with category information.
[0046]Drawing 14 is a figure showing other examples of a category estimation result display. 421 is the list of classification knowledge.
When the classification knowledge reference button 406 is pushed, it displays with reference to the classification knowledge 13.
422 is displayed with reference to the category definition table 14, when the category list button 405 is pushed. 423 is the text which described the range of a category.
In the category list 422, when any one category is chosen, it is displayed.

[0047]Drawing 15 is a figure showing an example of the screen after the log data was corrected by the user. About 411 and 412, a user can correct now the data displayed via the keyboard 20 and the mouse 25. In drawing 15, correction is made about 412. Deletion of the keyword currently displayed about the keyword, the addition of a new keyword, and correction of the dignity currently displayed are possible. To drawing 13 which is a screen before correction, by drawing 15, the dignity of keywords, such as a "circle", a "exchange market", and a "jump", is corrected, and the keyword which is not important is deleted out of "one day", "one time", etc.
[0048]Drawing 16 is a figure showing an example of a category re-estimation result. As a result of correcting a keyword and its dignity, it is shown that the category "money order" which did not appear as last estimation result as the classification result 414 appeared newly in primacy. Thus, about the category which newly appeared, the asterisk was added and it has distinguished from other categories. Of course, except addition of an asterisk may be sufficient as the method of distinction.
[0049]drawing 17 -- a category -- it is a figure showing the procedure of re-presumption 7. First, the table which stores the score of each category is initialized to 0 (Step 541).
[0050]Next, the document ID concerned, the text after correction, the keyword after correction, and its dignity are read in an output picture, and it stores in the working area 16 (Step 542). Next, it is distinguished whether text information was corrected (Step 543). If text information is corrected, in order for the keyword extracted from there and its dignity to change a lot, it is necessary to redo from the text analysis 2. since the keyword information read in the display screen can be used when text information is not corrected to it -- category presumption -- what is necessary is just to process from 3 A text correction flag is formed and it can be distinguished by the turning on and off whether text information was corrected.

[0051]At Step 543, when text information is corrected, text analysis 2 is performed, a keyword and dignity are extracted from the text after correction, and a result is stored in the working area 61 (Step 544).
[0052]Next, the following processings are performed about all the keywords stored in the working area 61 (Step 545). Distinguish whether the category in the classification knowledge containing the keyword concerned exists (Step 546), and about the existing category. The product of the dignity Wi (equivalent to 223 of drawing 6) which the keyword of the document concerned has, and the dignity Wj (equivalent to 244 of drawing 8) which the keyword of the category concerned has is calculated, and it adds as a score of the category concerned (Step 547).
[0053]Since it opts for the score of each category when all the keywords are followed, the deviation score of a score of each category is calculated from these scores (Step 548). A category is sorted in order with high deviation score (Step 549). And the document ID concerned, a category, and the value of the deviation score are made into a group, and it stores in the presumed category table 15 at order with high deviation score (Step 550).
[0054]Drawing 18 is a figure showing an example of the log data 17. In the log data 17, it stores and it is held until shut [the data about the items of the score according to the keyword extracted from the document ID and the text and its dignity, and category, and the fixed category]. Therefore, while checking the category estimation result of a certain document, the data of a checked document till then can also be referred to.
[0055]Drawing 19 is a figure showing an example of the category decision 8. A user becomes final and conclusive a category with reference to the classification result 414. In this example, the selected category is displayed on the definite category 415 by double-clicking a category to become final and conclusive with a mouse in the classification result 414.
[0056]Thus, according to this example, the candidate of a category is made to presume by a computer, the result is displayed and the man-machine assignment type document group supporting system that a user checks it can be realized to classify a document. Since it collects according to the presumed category and a result is shown one by one when displaying a classification result, the user can check efficiently. Even if the shown result is an error, data is corrected, by carrying out reclassification, the accuracy classified into a right category can be raised and the rate which does the big work of the burden that a user finds from 1 the category which should be classified can be lessened as much as possible.
[0057]Next, the modification of this example is described. In the similar document authorization 4, by this example, although recognized by top two presumed categories, it may recognize by a keyword with high dignity extracted from the text instead of being a presumed category.
[0058]Drawing 20 is a figure showing the disposal method. First, the similar document table 16 is initialized (Step 561). Next, the following processings are performed while the document which has not been authorized yet as a similar document exists (Step 562). The keyword beyond n kind (m>=n) of the keywords of m kind with dignity high about a certain document which is not authorized extracted from the document concerned, The document contained in the keyword of m kind with high dignity is extracted, and it stores in a similar document table with the set identifier for identifying a similar document set (Step 563). Drawing 11 defines a set identifier here as what substitutes it, although the name of the category was used as a thing

equivalent to a set identifier. As long as it is identifiable in a similar document set, what kind of form may be sufficient as this.

[0059]The document stored in the similar document table 16 is removed from the processing object of Step 562 after Step 563 (Step 564). It becomes possible to show a keyword with high dignity for every similar document [which is sharing] by the above processing, when showing a user the result by which category presumption was carried out.

[0060]Next, the extended example of this example is described. Like this example, when a category consists of two or more hierarchies, classification precision improvement can be expected by limiting to the subordinate category belonging to the superordinate category which shows a user a superordinate category, made specify it and was specified, and presuming a category. This is effective when the number of subordinate categories is huge especially.

[0061]Drawing 21 is a figure showing an example of the screen for specifying a large category. Specification of a large category is performed by displaying Screen 424 for specification, when the narrowing-down classification button 404 is pushed. Specification of a large category may be plural. Fundamentally, although the display order of the large category in Screen 424 for specification is an order defined as the category definition table 14, In adopting the technique of presuming a large category first and presuming a small category in category presumption 3 using the result, It is also possible to display the estimation result about the large category of the document concerned on the log data 17 by storing and holding based on an order of the estimation result of a large category.

[0062]the category limited to the large category specified by pushing the reclassification button 403 by Screen 424 for specification after specifying a large category -- re-presumption 7 is performed. the category shown in drawing 17 -- in Step 550 of the procedure of re-presumption 7, When an estimation result is stored in the presumed category table 15, it restricts, when the large category of the presumed category is contained in the large category specified by the user, and narrowing down by a superordinate category can be realized by storing. In the result display of drawing 13, temporarily, when a user narrows down a large category to "economy", in the classification result 414, the category the "political:Parliament" of the 2nd place is removed.

[0063]Thus, there are comparatively few superordinate categories, and when a user can become final and conclusive easily, a right category can be obtained by narrowing down by a superordinate category and presuming a category.

[0064]

[Effect of the Invention]Since a user is shown one by one for every document with which the result classified according to the computer was similar and a check is urged when a user checks the automatic-classification result of a document, It becomes easy to harness tips when a former document is checked, know how, teachings, data, etc. in the check of a next document, and check operation time ends few.

[0065]Since it is possible to draw a right classification result by making a user correct the log data outputted with an automatic-classification result, and re-presuming it even when an automatic-classification result is an error, When the first automatic-classification result is an error, the heavy work of the burden that a user reclassifies from 1 can be reduced.

[Translation done.]

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1]It is a figure showing the outline of this example.
[Drawing 2]It is a figure showing the outline of the hardware of this example.
[Drawing 3]It is a figure showing an example of the text contained in a document.
[Drawing 4]It is a figure showing an example of a word dictionary.
[Drawing 5]It is a figure showing an example of the word division results in text analysis.
[Drawing 6]It is a figure showing an example of a key word table.
[Drawing 7]It is a figure showing an example of a category definition table.
[Drawing 8]It is a figure showing an example of classification knowledge.
[Drawing 9]It is a figure showing the procedure of category presumption.
[Drawing 10]It is a figure showing an example of a presumed category table.
[Drawing 11]It is a figure showing the procedure of similar document authorization.
[Drawing 12]It is a figure showing an example of a similar document table.
[Drawing 13]It is a figure showing an example of a category estimation result display.
[Drawing 14]It is a figure showing other examples of a category estimation result display.
[Drawing 15]It is a figure showing an example of the screen after correction by a user.
[Drawing 16]It is a figure showing an example of a category re-estimation result.
[Drawing 17]It is a figure showing the procedure of category re-presumption.
[Drawing 18]It is a figure showing an example of log data.
[Drawing 19]It is a figure showing an example of category decision.
[Drawing 20]It is a figure showing other procedure of similar document authorization.
[Drawing 21]It is a figure showing an example of narrowing down of a superordinate category.
[Description of Notations]
A document input, 2:text analysis, 3:category presumption, 4 : 1: Similar document authorization, A category estimation result display, 6:log data correction, 7 : 5: Category re-presumption, 8: Category decision, 9:document data base registration, 10:document file, 11:word dictionary, 12:key word table, 13:classification knowledge, 14:category definition table, 15:presumption category table, 16:similar document table, 17:log data, 18 : document data base

[Translation done.]